Text Normalization Algorithm for Facebook Chats in Hausa Language

Jaafar Zubairu Maitama[1,2], Usman Haruna[1], Abdullahi Ya'u Gambo[1], Bimba Andrew Thomas[1], Norisma Binti Idris[1] Abdulsalam Ya'u Gital[3], Adamu I. Abubakar[4]

[1]Department of Artificial Intelligence, University of Malaya, Kuala Lumpur, Malaysia
[2]Faculty of Computer Science and Information Technology, Bayero University Kano
[3]Department of Mathematical Sciences, Abubakar Tafawa Balewa University, Bauchi Nigeria
[4]Department of Information Systems, International Islamic University Malaysia, Kuala Lumpur, Malaysia
jzzmaitama@gmail.com, [1]{agita,aiabubakar}1@acm.org

***Abstract*— The rapid increase in using non-standard words (NSWs) in communication through the social media is causing difficulties in understanding contents of the text messages. In addition, it affects the performance of several natural language processing (NLP) task such as machine translation, information retrievals, summarization and etc. In this study, we present an automatic text normalization system on Facebook chatting based on Hausa language. The proposed algorithm manually developed dictionary that employ normalization of each non-standard word with its equivalent standard word. This is accomplished through modification of the technique employed by [1] to fit Hausa NSWs' formation. It was found that our proposed algorithm was able to normalized Hausa NSWs with an accuracy of 100%. The results of this research can facilitate comprehensive communication via Facebook using Hausa language.**

Keywords: Hausa; Text Normalization; Non-standard word; Facebook Chat

I. INTRODUCTION

Non-standard words (NSWs) are words that are neither from dictionary or name, consisting of numbers, abbreviations, dates etc [2]. In this contemporary digital communication era, chatting via social media has become inevitable part of many people's daily activity and Facebook is viewed as the most foremost social network among youngsters [3]. However, the rapid increase in NSW usage in Facebook chats poses significant problems on content understanding and they are ambiguous.

The rationale behind text normalization involves establishing a standard rule for transforming text into a single standard form that it might not have had before. Text normalization is necessary due to the fact that it helps in given the complete and formal version of text abnormal appearance. This helps in conveying the right information about the content of a chat. In addition, its support further activities of natural Language processing (NLP) on the charts. We observed that, lack of text normalization causes unsuccessful Facebook chat communication due to difficulties in content understanding by users.

In an effort to circumvent the prevailing problems associated with restoration of the correct word in social media Liu *et al*. [4] introduce the use of human perspectives approach namely, enhancement letter transformation, visual priming and phonetic similarity. Similarly, some researchers suggest that the problem can be addressed using a phrase based statistical model [5], an unsupervised approach [6, 7], letter level alignment [7] or combining automated and manual method [1]. These techniques effectively normalized NSWs to their standard form with alternating results, but they are only available in a limited number of languages which include English language.

There are, however, to the best of our knowledge, lack of studies that attempt to perform standard words (SWs) of Hausa language despite Hausa language is an international language widely spoken in West Africa, with an estimated spoken populace of 52 million. It is a very popular language in many African countries like Nigeria, Niger, Cameroon, Benin, Togo, Ghana and etc. in which Nigeria is the most populous in the entire Africa. With regards to peoples' mother tongue, Hausa happens is the first among all the languages in Africa [8]. Like any other language, it has its own system of writing such as Ajami and Boko. Ajami makes use of Arabic alphabets whereas Boko utilize the Romance alphabets [9]. Nigeria is the country with the highest number of Hausa speakers [8], Facebook appeared to be the most populous social media platform for communication using Hausa language with over 4 million users as of October 2011 with 300% increment from 2010 [10]. In this paper, we proposed to modify the algorithm originally propose by [1] to identify and normalize Hausa NSWs to their equivalent standard words (SWs).

The rest of the paper is organized as follows:. Section II introduces the proposed method with its implementation details. Section III presents results and discussion before concluding remarks and further research direction in section IV.

II. METHODOLOGY

The basic steps in the design of the algorithm for the text normalization in Hausa language involve the following steps:

A. Identification of the non-standard Hausa words on Facebook chats

Identification of the non-standard Hausa words on Facebook chats: The initial stage begins by conducting an experiment with the aim of identifying Hausa non-standard words and how they are formed. Samples of charts were collected from Facebook accounts of 30 different Hausa speakers. Document analysis method was

used to analyse the chats consisting of 806 words and gather qualitative information associated with the chat production. To ensure reliability as well as unbiased experiment, the sample data were grouped into 3 files, namely; *chat, chat2, and chat3*. The first sample was the use of the dictionary development, while the remaining was purely involved in the system and the evaluation. It was found that Hausa non-standard words are generally produced using an abbreviation or vowel omission, with a few omissions of both vowels and consonants in rare cases. This is achieved by simply deleting one or more vowel or combination of consonants and vowels of a particular word to generate a non-standard words, see Table I.

TABLE I. SAMPLES OF WORDS USED

S/N.	Original word	Non-standard word	Character Omitted	Type of character
1.	Tafiya	Tfya	a, i	Vowel
2.	Gida	Gda	I	Vowel
3.	Ranar	Rnr	a, a	Vowel
4.	Talata	Tlt	a, a, a	Vowel
5	Cancanta	Ccnta	a, n, a	Both
5.	Wannan	Wnn	a, n, a	Both
6.	Wallahi	Wlhi	a, l, a	Both

B. *Dictionary development*

Non-standard words identified from the first sample *chat* with 278 words were used to build up a dictionary. The dictionary is a database that consists of a set of non-standard words mapped with their equivalent standard words. It can be updated with a newly found non-standard word which does not exist. This unique feature enables the dictionary to handle as many as possible NSWs for effective normalization. The database is developed in the Mysql database platform.

C. *System Overview*

An overview of the system is required to be introduce similar to the studies in [11-13]. The flow of our proposed algorithm is presented schematically in Fig. 1 which gives the full reprentation of the system.

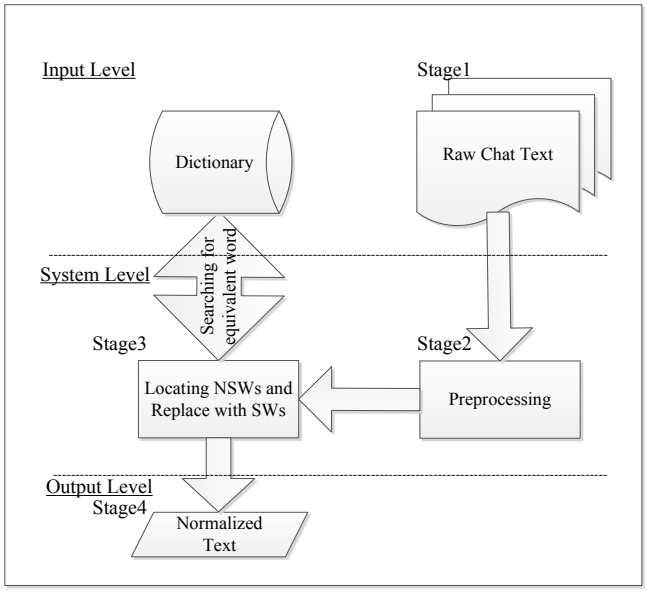

Figure 1. System Architecture

The system was developed using Python programming language. It begins by reading the Hausa Facebook chats inputs from the collected chat samples saved in notepad. One file of a sample chats uploads at a time. Subsequently, will undergoes preprocessing, which involves removal of punctuations, changing the entire chat case to lower case as well as tokenizing the chats. The dictionary lookup is performed from the developed dictionary and finally return the normalized version of the chat. The dictionary lookup is performed in such a way that the system can search for the existence of non-standard word from the inputs chat, if found then the system replace it with its equivalent SWs in the dictionary.

In view of the fact that Hausa and English languages share one major NSW categorization pattern called "*Abbreviation*" and based on our dictionary approach [1] is the most effective means to normalize NSWs produced using *abbreviations*. However, as this adopted approach involves the use of motifs method which discovered to be inapplicable in Hausa NSWs. Thus, this made it obligatory to us to modify the algorithm so as to fit Hausa NSWs' formation and hence became more special and useful compared to other techniques.

D. *The Proposed Hausa Normalization Algorithm*

The summary of the proposed algorithm of the system is as follows:

Step1: start;
Step2: Read the raw data;
Step3: Tokenize the data;
Step4: Search for a word;
Step5: if the word searched is SW then goto step 7;
Step6: else Replace the word with SW
Step7: if end of input-data is not reached then goto step 4;
Step8: else output Normalized data
Step9: end;

III. Result And Discussion

The system was evaluated by the 3 sample chats collected. Experimental results of the chats are observed and recorded. Chat2 and Chat3 are involved at this stage. These chat samples were read one at a time, followed by their respective analysis. After the algorithm run for all the three files, the results were recorded and presented in Table II.

TABLE II. Experimental results of the chats

Sample Chats	Chat	Chat2	Chat3	Chat4
No. of non-normalize NSW	0	55	130	121
No. of NSW	129	102	173	699
No. of normalized NSW	229	47	43	578
No. of input word recorded to system	278	173	355	1,023
Recall in (%)	46.4	58.9	48.7	68.3
Accuracy in (%)	100	46.1	24.9	82.7

The results show that out of 129 NSWs found in 'chat' file, that contain 278 words, the system was able to normalize all the 129 NSW. It achieved 100% accuracy. Probably the results were achieved because the NSW were contained in the diction which could have made it possible for the algorithm to correctly identify the equivalent SWs. This indicates that all the NSWs were identified and normalized correctly. On the other hand, caht2 attain 46.1% accuracy whereby it was able to normalize relatively half of the entire NSWs identified. This result performed poorly compared to the chat file results, probably the results performs poorly due to the fact that none of its words are used in developing the dictionary. Thus, some newly NSWs are introduced that need to be included in the database.

However, chat3 yield the list accuracy simply because it is the chat sample that contains the highest number of NSWs and only a few are in the dictionary. This causes significantly low NSWs identification and eventually leads to inability to normalize such words. Finally a 1,023 chats were collected, named *Cha4* and input to the system for NSWs' normalization. The system achieved an impressive result of 82.7% accuracy more than twice that of *chat3*. Although, it was not 100% as seen in chat because some of the NSWs were not in the dictionary, but appeared to be effective. The performance is perhaps due to continued updates of the dictionary with newly discovered NSWs. After we obtained results of *chat3*, we then update the dictionary some NSWs that were not part of the dictionary earlier. The majority of which were found in *chat3*.

IV. Conclusions And Further Work

The system propose in the study can effectively identify and normalized Hausa NSWs to SWs with an accuracy of 100%. This is evident from the results obtained, even though it shows to perform less for the other three sets of data. Still, it performances remain appreciable when compared to existing techniques, whereby it is believed to have better accuracy in the presence of a well saturated dictionary. To enhance its effectiveness in normalization of any set of data, large data set of NSW is required to build up the dictionary. In this paper no any corpus is used, and NWS was counted manually for determination of the system accuracy. This is prone to error due to human nature. Future work will include corpus that can be used to build up the system and the system should be made open to users for regular update of newly introduced NSWs. Also, we intend to used soft computing approaches [14-16] in dealing with the NSWs.

References

[1] E. Clark and K. Araki, "Text normalization in social media: progress, problems and applications for a pre-processing system of casual English," *Procedia-Social and Behavioral Sciences,* vol. 27, pp. 2-11, 2011.

[2] R. Sproat, A. W. Black, S. Chen, S. Kumar, M. Ostendorf, and C. Richards, "Normalization of non-standard words," *Computer Speech & Language,* vol. 15, pp. 287-333, 2001.

[3] C.-c. Yang and B. B. Brown, "Motives for using facebook, patterns of facebook activities, and late adolescents' social adjustment to college," *Journal of youth and adolescence,* vol. 42, pp. 403-416, 2013.

[4] F. Liu, F. Weng, and X. Jiang, "A broad-coverage normalization system for social media language," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, 2012, pp. 1035-1044.

[5] A. Aw, M. Zhang, J. Xiao, and J. Su, "A phrase-based statistical model for SMS text normalization," in *Proceedings of the COLING/ACL on Main conference poster sessions*, 2006, pp. 33-40.

[6] P. Cook and S. Stevenson, "An unsupervised model for text message normalization," in *Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*, 2009, pp. 71-78.

[7] F. Liu, F. Weng, B. Wang, and Y. Liu, "Insertion, deletion, or substitution?: normalizing text messages without pre-categorization nor supervision," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, 2011, pp. 71-76.

[8] M. A. Smirnova, *The Hausa language: a descriptive grammar*: Routledge & Kegan Paul, 1982.

[9] R. M. Newman and P. Newman, "The Hausa lexicographic tradition: lexikovaria," *Lexikos,* vol. 11, pp. p. 263-286, 2001.

[10] C. Fink, J. Kopecky, N. Bos, and M. Thomas, "Mapping the Twitterverse in the developing world: An analysis of social media use in Nigeria," in *Social Computing, Behavioral-Cultural Modeling and Prediction*, ed: Springer, 2012, pp. 164-171.

[11] T. Mantoro, A.I. Abubakar, and H. Chiroma, 'Pedestrian position and pathway in the design of 3D mobile interactive navigation aid', in Editor (Ed.)^(Eds.): 'Book Pedestrian position and pathway in the design of 3D mobile interactive navigation aid' (ACM, 2012, edn.), pp. 189-198

[12] J. Memon, A.R.M. Zaidi, M. Uddin, M., A. I. Abubakar, H. Chiroma, and D. Daud, 'Randomized Text Encryption: a New Dimension in Cryptography', International Review on Computers and Software (IRECOS), 2014, 9, (2)

[13] Zeki, A.M., Elnour, E.E., Ibrahim, A.A., Haruna, C., and Abdulkareem, S.: 'Automatic interactive security monitoring system', in Editor (Ed.)^(Eds.): 'Book Automatic interactive security monitoring system' (IEEE, 2013, edn.), pp. 215-220

[14] Abubakar, A., Zeki, A.M., and Chiroma, H.: 'Optimizing Three-Dimensional (3D) Map View on Mobile Devices as Navigation Aids Using Artificial Neural Network', in Editor (Ed.)^(Eds.): 'Book Optimizing Three-Dimensional (3D) Map View on Mobile Devices as Navigation Aids Using Artificial Neural Network' (IEEE, 2013, edn.), pp. 232-237

[15] A.I. Abubakar, A.M. Zeki, H. Chiroma, T. Herawan, 'Investigating Rendering Speed and Download Rate of Three-Dimension (3D) Mobile Map Intended for Navigation Aid Using Genetic Algorithm': 'Recent Advances on Soft Computing and Data Mining' (Springer, 2014), pp. 261-271

[16] H. Chiroma, A. Gital, A. I. Abubakar, A.M. Zeki, 'Comparing Performances of Markov Blanket and Tree Augmented Naïve-Bayes on the IRIS Dataset', in Editor (Ed.)^(Eds.): 'Book Comparing Performances of Markov Blanket and Tree Augmented Naïve-Bayes on the IRIS Dataset' (2014, edn.), pp.